Bayesian Causal Inference - Session 1

Bayesian Regression and Computation

Arman Oganisian
Thomas J. & Alice M. Tisch Assistant Professor of Biostatistics

Department of Biostatistics
Brown University

5/29/2025

Outline

Session 1: Bayesian Crash Course

- Priors, Posteriors, Likelihoods.
- Posterior Computation.
- Shrinkage and Comparison with frequentist methods.

Session 2: Bayesian Causal Inference

- Hierarchical Causal Inference.
- Sensitivity Analysis.
- Bayesian ML methods.

Prerequisites and Objectives

By the end of Session 1, I am hoping you will:

- Be able to describe the differences between Bayesian and frequentist inference.
- Appreciate the main selling points of the Bayesian approach.
- Have awareness of statistical software for Bayesian computation `R`.

I will assume the following knowlege throughout:

- A graduate-level course in probability theory.
- A graduate-level course in statistical inference.
- Familiarity with statistical computing in `R`.
- Previous summer institute sessions (in particular, Peter Yang's session on Day 1).

Popular Examples of Bayesian Methods

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 31, 2020 VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D.,
Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D.,
Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D.,
Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D.,
Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D.,
Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D.,
Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

ABSTRACT

Popular Examples of Bayesian Methods

The NEW ENGLAND JOURNAL of MEDICINE

laboratory or at a local testing facility (using a protocol-defined acceptable test).

Major secondary end points included the efficacy of BNT162b2 against severe Covid-19. Severe Covid-19 is defined by the FDA as confirmed Covid-19 with one of the following additional features: clinical signs at rest that are indicative of severe systemic illness; respiratory failure; evidence of shock; significant acute renal, hepatic, or neurologic dysfunction; admission to an intensive care unit; or death. Details are provided in the protocol.

tions (the population that could be evaluated). Vaccine efficacy was estimated by $100 \times (1 - \text{IRR})$, where IRR is the calculated ratio of confirmed cases of Covid-19 illness per 1000 person-years of follow-up in the active vaccine group to the corresponding illness rate in the placebo group. The 95.0% credible interval for vaccine efficacy and the probability of vaccine efficacy greater than 30% were calculated with the use of a Bayesian beta-binomial model. The final analysis uses a success boundary of 98.6% for probability of vaccine efficacy greater than 30% to

Popular Examples of Bayesian Methods

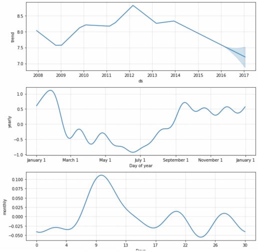
PROPHET Docs GitHub

The inputs to this function are a name, the period of the seasonality in days, and the Fourier order for the seasonality. For reference, by default Prophet uses a Fourier order of 3 for weekly seasonality and 10 for yearly seasonality. An optional input to `add_seasonality` is the prior scale for that seasonal component - this is discussed below.

As an example, here we fit the Peyton Manning data from the Quickstart, but replace the weekly seasonality with monthly seasonality. The monthly seasonality then will appear in the components plot:

```r
# R
m <- prophet(weekly.seasonality=FALSE)
m <- add_seasonality(m, name='monthly', period=30.5, fourier.order=5)
m <- fit.prophet(m, df)
forecast <- predict(m, future)
prophet_plot_components(m, forecast)
```

```python
# Python
m = Prophet(weekly_seasonality=False)
m.add_seasonality(name='monthly', period=30.5, fourier_order=5)
forecast = m.fit(df).predict(future)
fig = m.plot_components(forecast)
```



… that could be evaluated). …estimated by 100 × (1 − IRR), …culated ratio of confirmed …ness per 1000 person-years …ctive vaccine group to the … rate in the placebo group. …nterval for vaccine efficacy …of vaccine efficacy greater …ulated with the use of a …al model. The final analy- …undary of 98.6% for prob- …icacy greater than 30% to

Popular Examples of Bayesian Methods

Uber Schedule rides in advance Reserve a ride ×

Data / ML, Engineering

Introducing Orbit, An Open Source Package for Time Series Inference and Forecasting

May 14, 2021 / Global

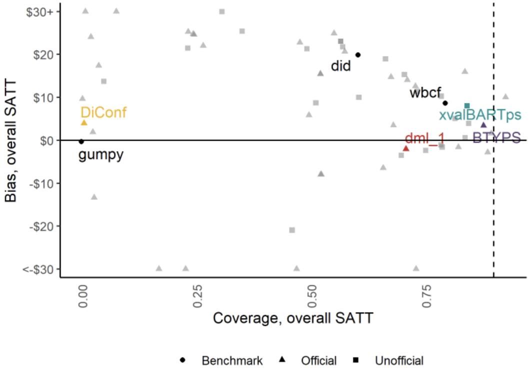
Orbit

Orbit is a general interface for Bayesian time series modeling. The goal of Orbit development team is to create a tool that is easy to use, flexible, interitible, and high performing (fast computation). Under the hood, Orbit uses the probabilistic programming languages (PPL) including but not limited to Stan and Pyro for posterior approximation (i.e, MCMC sampling, SVI). Below is a quadrant chart to position a few time series related packages in our assessment in terms of flexibility and completeness. Orbit is the only tool that allows for easy model specification and analysis while not limiting itself to a small subset of models. For example Prophet has a complete end to end solution but only has one model type and Pyro has total specification model flexibility but does not give an end to end solution. Thus Orbit bridges the gap between business problems and statistical solutions.

Popular Examples of Bayesian Methods

Schedule rides

Figure 1: For the overall SATT, few estimators minimized confounding bias while achieving nominal coverage.



Note: The vertical dashed line shows nominal coverage of the uncertainty intervals (0.9).

specification model flexibility but does not give an end to end solution. Thus Orbit bridges the gap between business problems and statistical solutions.

Review of Frequentist Inference

Defining Notation

Consider binary outcome data on n subjects, $\boldsymbol{Y}_n = (Y_1, Y_2, \ldots, Y_n)$

$$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} f_{Y|\omega}(y \mid \omega)$$

- Y_i: a observed outcome for subject i that takes values in $\mathcal{Y}$.
- $f_{Y|\omega}(y \mid \omega)$: density/mass function function for each $y \in \mathcal{Y}$.
- ω: a real-valued parameter that takes values in $\mathcal{P}$.
- *iid* indicates these are "independent and identically distributed" outcome data.
- Will use uppercase to denote random variable and lowercase to denote realization (e.g. $\boldsymbol{Y}_n$ versus $\boldsymbol{y}_n$).

Defining Notation - Example

Consider binary outcome data on $n = 30$ subjects, $\boldsymbol{Y}_n = (Y_1, Y_2, \ldots, Y_{30})$

$$Y_1, Y_2, \ldots, Y_{30} \overset{iid}{\sim} Ber(\omega)$$

- Y_i: a observed outcome for subject i that takes values in $\mathcal{Y} = \{0, 1\}$.
- $Ber(\omega)$ indicates the Bernoulli distribution with pmf $f_{Y|\omega}(y \mid \omega) = \omega^y (1 - \omega)^{1-y}$ for $y \in \{0, 1\}$.
- ω: a parameter representing the probability $P(Y = 1) = \omega$, takes values in $\mathcal{P} = [0, 1]$.

Probability in Frequentist Inference

$$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} f_{Y|\omega}(y \mid \omega)$$

In frequentist paradigm,

- Parameter ω considered fixed, unknown truth in nature.
- Data, $\boldsymbol{Y_n}$, are considered random variables: each hypothetical re-sampling yields a different $\boldsymbol{Y}_n$.
- The distribution $f_{Y|\omega}$ characterizes behavior of $\boldsymbol{Y_n}$ across repeated hypothetical re-samplings - i.e. sampling variability.
- Probability of an event A, $P_{Y|\omega}(Y \in A \mid \omega)$, represents the long-run "frequency" of the event $Y \in A$ occuring across repeated hypothetical re-samplings of $\boldsymbol{Y}_n$.
- Frequentist statistics finds methods of making inferences on ω that have "good" repeated sample (i.e. frequentist) properties.

Point Estimation

A point estimator $\hat{\omega}(\boldsymbol{Y}_n)$ for ω is a mapping from data to the parameter space, $\hat{\omega}(\boldsymbol{Y}_n) : \boldsymbol{\mathcal{Y}}_n \rightarrow \mathcal{P}$.

Since the data are random, $\hat{\omega}$ is random.

It also has long-run operating characteristics:

- Bias: $\text{Bias}(\hat{\omega}) = \text{E}_{Y|\omega}[\hat{\omega}(\boldsymbol{Y}_n)] - \omega$.
- Variance: $\text{V}(\hat{\omega}) = \text{E}_{Y|\omega}\left[\left(\hat{\omega}(\boldsymbol{Y}_n) - \text{E}_{Y|\omega}[\hat{\omega}(\boldsymbol{Y}_n)]\right)^2\right]$.
- Mean squared-error: $\text{MSE}(\hat{\omega}) = \text{E}_{Y|\omega}[(\hat{\omega} - \omega)^2] = \text{Bias}(\hat{\omega})^2 + \text{V}(\hat{\omega})$

In addition, it has modes of convergence as $n \rightarrow \infty$.

Interval Estimation

For some $\alpha \in (0, 1)$, a 100$(1-\alpha)$% confidence interval is given by $(L(\boldsymbol{Y}_n), U(\boldsymbol{Y}_n))$ such that $L(\boldsymbol{Y}_n) < U(\boldsymbol{Y}_n)$ and

$$P_{Y|\omega}\Big(L(\boldsymbol{Y}_n) < \omega < U(\boldsymbol{Y}_n)\Big) = 1 - \alpha$$

Since the data are random, the interval end points are random.

- "Confidence" refers to the long-run proportion of times that an interval estimator would contain the true parameter value, ω.
- Intervals have long-run operating characteristics too. E.g. expected width, $E_{Y|\omega}[U(\boldsymbol{Y}_n) - L(\boldsymbol{Y}_n)]$.

Hypothesis Testing

Consider testing $H_0 : \omega = \omega_0$ against $H_a : \omega \neq \omega_0$. A hypothesis test, R, is a mapping from data to a (random) decision: $R : \boldsymbol{\mathcal{Y}}_n \to \{0, 1\}$, where $R(\boldsymbol{Y}_n) = 1$ indicates a rejection of H_0 and $R(\boldsymbol{Y}_n) = 0$ indicates failure to reject.

Since the data are random, the decision is random.
The test R is constructed to have specified repeated sample properties:

- Type 1 error, α: proportion of times we can expect to reject H_0 if it were true.
- Type 2 error: proportion of times we will fail to reject H_0 if H_a were true.

Different Estimators Have Different Properties

"There are no solutions, only trade-offs" - Thomas Sowell

Making inferences involves trade-offs:

- Bias-Variance tradeoff: Consider point-estimator, $\hat{\omega}(\boldsymbol{Y}_n) = 1$. This estimator has zero variance. But it is biased for all $\omega \neq 1$.
- Width - Confidence Level tradeoff. The confidence interval $[L(\boldsymbol{Y}_n) = 0, U(\boldsymbol{Y}_n) = \infty]$ for average height in the city of Providence (in centimeters) has 100% confidence level. However, it is much too wide to be useful.
- Type 1 and Type 2 error tradeoff. Consider a test that *always fails to reject* the null. This test has 0% Type 1 error. However, the Type 2 error rate of this test would be 100% since the null would always be retained, even if the alternative were true.

An Example: Estimating a Binomial Proportion

Suppose $Y_1, Y_2, \ldots, Y_n \sim Ber(\omega)$.

- An unbiased point estimator is given by Maximum Likelihood Estimator (MLE):

$$\hat{\omega} = \bar{y}_n = \text{argmax}_{\omega \in \mathcal{P}} \mathcal{L}(\omega \mid \boldsymbol{y}_n)$$

  where $\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\mathcal{L}(\omega \mid \boldsymbol{y}_n) = \prod_{i=1}^{n} f_{Y|\omega}(y_i \mid \omega)$ is the likelihood.

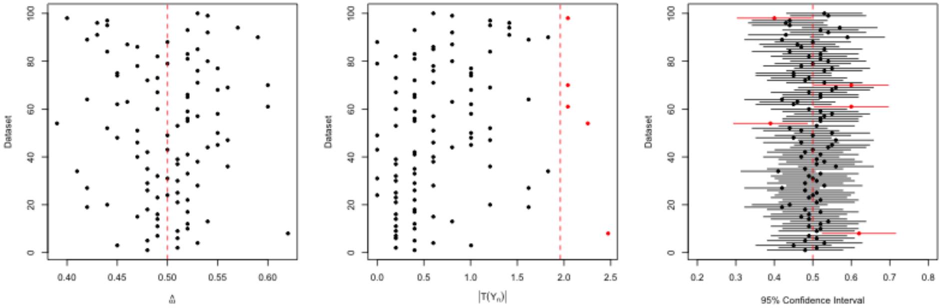
- An $\alpha = .05$ level test of $H_0 : \omega = .5$ is given by $R(\boldsymbol{Y}_n) = I(|T(\boldsymbol{Y}_n)| > z_{.975})$ where under the null

$$T(\boldsymbol{Y}_n) = \frac{\sqrt{n}(\hat{\omega} - .5)}{\sqrt{\hat{\omega}(1 - \hat{\omega})}} \overset{A}{\sim} N(0, 1)$$

- A 95% confidence interval can be found by inverting the test above.

$$\bar{y}_n \pm z_{.975} \sqrt{\frac{\bar{y}_n(1 - \bar{y}_n)}{n}}$$

Inference for Binomial Proportion



see 1_freq_examples.R

Are frequentist solutions satisfying?

Suppose we compute interval $[-10.23, 5.21]$.

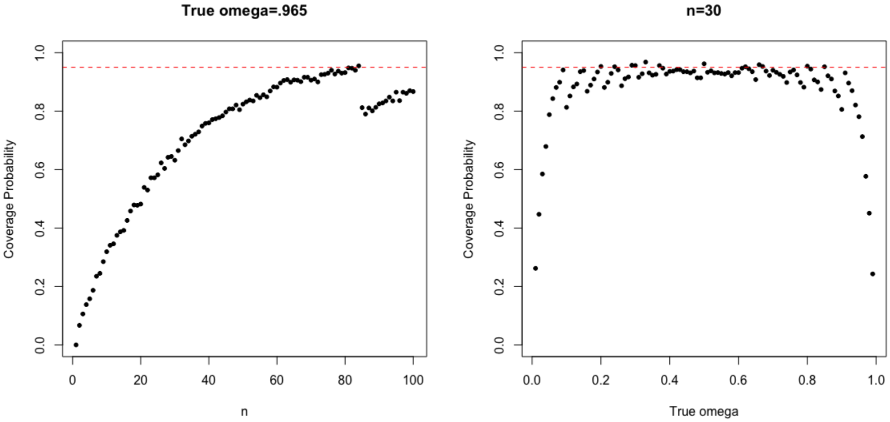
- What we want to say: "There is a 95% probability that ω is between -10.23 and 5.21."
- Wrong: ω is fixed, not a RV. It's either in $[-10.23, 5.21]$ or it isn't.
- What we can say: "95% of intervals constructed in this way across hypothetical resamplings of the data will contain ω.

Suppose we compute a p-value of .023

- What we want to say: "There is a 2.3% probability that the null hypothesis is true."
- Wrong: Either $\omega = \omega_0$ or it doesn't. There's no probability associated with this since it is a constant, not a RV.
- What we can say: "Assuming the null is true, there is only a 2.3% chance of observing a test statistic as or more extreme than the one observed."

Are frequentist solutions satisfying?

Frequentist gaurantees hold on average across hypothetical repitions, as sample size goes to infinity. In practice: we only have one study with finite n that cannot be repeated.



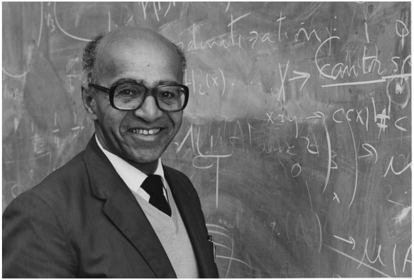
see 2_bin_interval.R

Blackwell on Bayesian Inference

"... An economist came in one day to talk to me while I was visiting there. And he said, 'I need a number. I need to know the probability of a major war within the next five years.' And he explained to me why he needed to know that number and it made a lot of sense...I said, 'The concept of probability makes sense only in a long sequence of events under identical conditions.' And the occurrence of a war in the next five years is a unique phenomenon and the probability is either zero or one and we won't know for five years. And he looked at me, and he said, 'Thank you.' He said that he had spoken with several other statisticians and they'd all told him the same thing, and he left.

That conversation bothered me. The man had asked me a serious, reasonable question and I had given him a kind of flip answer..."

David Blackwell



Crash Course on Bayesian Inference

Probability: The Bayesian Perspective

Probability is *not* long-run frequency of events in repeated hypothetical resamplings - but a measure of uncertainty.

1. Uncertainty about possible data realizations $\rightarrow$ probability distributions for $\boldsymbol{Y}_n$.
2. Uncertainty about model parameter $\rightarrow$ probability distributions for ω.

Since we have uncertainty about both, both are RVs with their own probability distributions. Can make probability statements about both before and after having considered the data.

The Bayesian Perspective

Notation: $\mathcal{P}$ will now denote the parameter space, $\omega \in \mathcal{P} \subset \mathbb{R}$. We will use uppercase Ω denote the random variable and lowercase ω to denote a realization $\Omega = \omega$. For now, $\dim(\mathcal{P}) = 1$.

1. Given some ω, the model $f_{\boldsymbol{Y}_n|\Omega}(\boldsymbol{y}_n \mid \omega)$ describes uncertainty about data.
2. Prior Distribution: the model $f_{\Omega}(\omega)$ describes *initial* state of uncertainty about ω, *before considering the data*. Note: usually has hyperparameters γ, so should be $f_{\Omega|\Gamma}(\omega \mid \gamma)$.
3. Posterior Distribution: $f_{\Omega|\boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n)$ describes *revised/updated* state of uncertainty about ω, *after considering the data*.

We will sometimes omit the subscripts on f when it is obvious which density/mass functions are being referenced.

Updating Uncertainty via Bayes' Rule

Bayesian inference - in all settings - follows a unified procedure: find a distribution over quantities you want to know, conditional on the quantities you do know.

$$
\begin{aligned}
f_{\Omega \mid \boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n) &= \underbrace{\frac{1}{f_{\boldsymbol{Y}_n}(\boldsymbol{y}_n)}}_{\text{Normalizing Constant}} \cdot \underbrace{\mathcal{L}(\omega \mid \boldsymbol{y}_n)\ f_{\Omega}(\omega)}_{\text{Unnormalized Posterior}} \\
&\propto \mathcal{L}(\omega \mid \boldsymbol{y}_n)\ f_{\Omega}(\omega)
\end{aligned}
$$

- Evidence: $f_{\boldsymbol{Y}_n}(\boldsymbol{y}_n) = \int_{\mathcal{P}} f_{\boldsymbol{Y}_n \mid \omega}(\boldsymbol{y}_n \mid \omega)\ f_{\Omega}(\omega) d\omega$ is typically unknown.
- Likelihood: Under *iid* sampling, $\mathcal{L}(\omega \mid \boldsymbol{y}_n) = \prod_{i=1}^{n} f_{Y \mid \omega}(y_i \mid \omega)$

Updating Uncertainty via Bayes' Rule

In the Bayesian inferential paradigm, all inference is based on posterior. Common summaries are integrals over the posterior:

- Posterior Point Estimation:

$$E_{\Omega|\boldsymbol{Y}_n}[\Omega \mid \boldsymbol{y}_n] = \int_{\mathcal{P}} \omega f_{\Omega|\boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n) d\omega$$

  .

- $100(1-\alpha)\%$ Credible Set Estimation: Find region $C(1-\alpha) \subset \mathcal{P}$ such that

$$\int_{C(1-\alpha)} f_{\Omega|\boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n) d\omega = 1 - \alpha$$

Finding the Posterior

$$f_{\Omega|\boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n) \propto \mathcal{L}(\omega \mid \boldsymbol{y}_n)\ f_\Omega(\omega)$$

The main task of Bayesian inference is finding the posterior disrtribution. Given a likelihood of your model and the prior density, there are two strategies:

1. Analytic: Solve for the evidence. The normalize by the inverse of the evidence to get $f_{\Omega|\boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n)$.
2. Computational: Using Markov Chain Monte Carlo (MCMC) methods simulate a set of draws from $f_{\Omega|\boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n)$

We will focus on computational strategies in this course.

Inference using Posterior Draws

Suppose we have a set of draws $\{\omega^{(m)}\}_{m=1}^{M}$ from the posterior. Then, posterior summaries can be constructed by post-processing these draws

- Posterior Point Estimation:

$$E_{\Omega \mid \boldsymbol{Y}_n}[\Omega \mid \boldsymbol{y}_n] = \frac{1}{M}\sum_{m=1}^{M}\omega^{(m)}$$

- $100(1-\alpha)\%$ Credible Set Estimation: Find region $C(1-\alpha) \subset \mathcal{P}$ such that

$$\int_{C(1-\alpha)} f_{\Omega \mid \boldsymbol{Y}_n}(\omega \mid \boldsymbol{y}_n)\,d\omega = 1-\alpha$$

  Take $C(1-\alpha)$ to be the interval between the $(\alpha/2)^{th}$ and $((1-\alpha)/2)^{th}$ percentiles of $\{\omega^{(m)}\}_{m=1}^{M}$.

Inference for Transformations

Posterior summaries for transformations $\Theta = g(\Omega)$ can be constructed by post-processing these draws

$$f_{\Theta|\boldsymbol{Y}_n}(\theta \mid \boldsymbol{y}_n) = f_{\Omega|\boldsymbol{Y}_n}(g^{-1}(\theta) \mid \boldsymbol{y}_n)\left|\frac{d}{d\theta}g^{-1}(\theta)\right|$$

- Posterior Point Estimation:

$$E_{\Theta|\boldsymbol{Y}_n}[\Theta \mid \boldsymbol{y}_n] = \frac{1}{M}\sum_{m=1}^{M} g(\omega^{(m)})$$

- $100(1-\alpha)\%$ Credible Set Estimation: Take $C(1-\alpha)$ to be the interval between the $(\alpha/2)^{th}$ and $((1-\alpha)/2)^{th}$ percentiles of $\{g(\omega^{(m)})\}_{m=1}^{M}$.

Posterior for Binomial Proportion

Bayesian Inference for Binomial Proportion

Suppose we have binary outcome $Y_i \in \{0, 1\}$ and prior $\Omega \sim Beta(\alpha, \beta)$

$$Y_1, Y_2, \ldots, Y_n \mid \Omega \overset{iid}{\sim} Ber(\omega)$$
$$\Omega \sim Beta(\alpha, \beta)$$

- The Beta density is $f_\Omega(\omega) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\omega^{\alpha-1}(1-\omega)^{\beta-1}$ for $\omega \in (0, 1)$
- It is a convenient choice due to proper support on values $0 \leq \omega \leq 1$.
- Prior mean is $E[\Omega] = \frac{\alpha}{\alpha+\beta}$.
- Posterior can be found analytically to be another Beta:

$$\Omega \mid \boldsymbol{Y}_n \sim Beta(n\bar{y}_n + \alpha, n(1 - \bar{y}_n) + \beta)$$

where $\bar{y}_n = (1/n)\sum_{i=1}^{n} y_i$.

Bayesian Inference for Binomial Proportion

$$\Omega \mid \boldsymbol{Y}_n \sim Beta(n\bar{y}_n + \alpha, n(1-\bar{y}_n) + \beta)$$

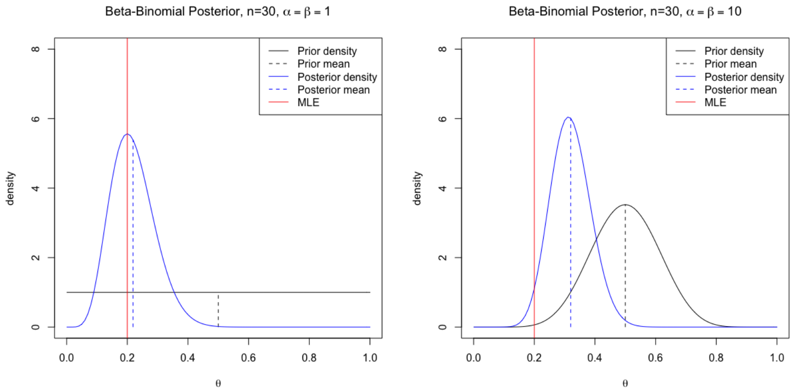
where $\bar{y}_n = (1/n)\sum_{i=1}^{n} y_i$.

The posterior mean is given by

$$E[\Omega \mid \boldsymbol{Y}_n] = \frac{n}{n+\alpha+\beta}\bar{y}_n + \frac{\alpha+\beta}{n+\alpha+\beta} \cdot \frac{\alpha}{\alpha+\beta}$$

- A weighted average of prior mean $E[\Omega] = \frac{\alpha}{\alpha+\beta}$ and MLE $\bar{y}_n$.
- For small n, MLE is shrunk towards prior mean, $E[\Omega \mid \boldsymbol{Y}_n]$ puts weight nearly 1 on $\frac{\alpha}{\alpha+\beta}$.
- As $n \to \infty$, posterior mean $E[\Omega \mid \boldsymbol{Y}_n]$ puts weight 1 on $\bar{y}_n$.

Informative and Uninformative Priors



- Informative priors: prior distributions that are more impactful on the posterior. aka: "tight" priors.
- Uninformative priors: prior distributions that are less impactful on the posterior. aka: "flat" priors or "wide priors"

Computation via Simulation

Since we know

$$\Omega \mid \boldsymbol{Y}_n \sim Beta(n\bar{y}_n + \alpha, n(1-\bar{y}_n) + \beta)$$

we can obtain draws $\{\omega^{(m)}\}_{m=1}^{M}$ easily. E.g, for $M = 100,000$,

```
# simulate 100,000 values from the posterior
post_draws = rbeta(100000, n*y_bar+alpha, n*(1-y_bar)+beta )

## approximate posterior mean
mean( post_draws )

## find 95\% credible interval
quantile( post_draws , probs = c( .025 , .975 ) )
```

Inference for Transformations

What if we want to make inferences about $\Theta = g(\Omega) = \frac{\Omega}{1-\Omega}$?
Can be done via simulation:

```
# simulate 100,000 values from the posterior of omega
post_draws = rbeta(100000, n*y_bar+alpha, n*(1-y_bar)+beta )

## compute odds
post_draws_odds = post_draws / ( 1 - post_draws )

mean( post_draws_odds )
quantile( post_draws_odds , probs = c( .025, .5, .975 ) )
```

Priors as Shrinkage

Recall the previously discussed frequentist interval,

$$\bar{y}_n \pm z_{.975}\sqrt{\frac{\bar{y}_n(1-\bar{y}_n)}{n}}$$

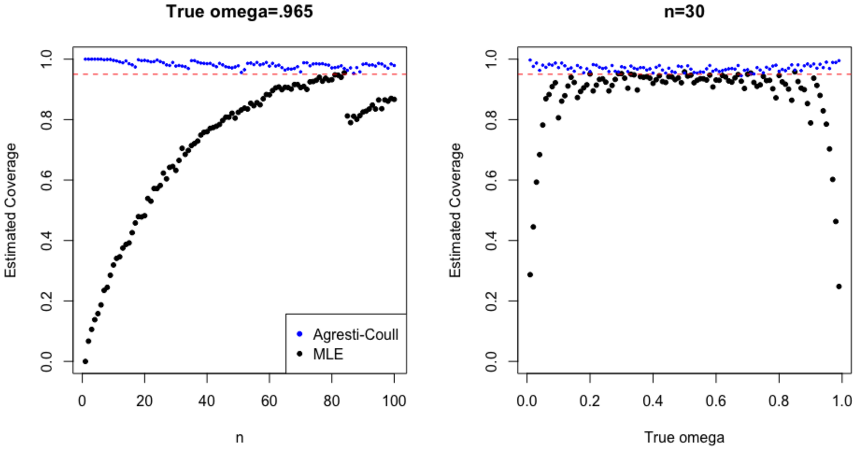
, performed poorly at the edge of parameter space and for small n.
One alternative is is the Agresti-Coull interval estimator:

$$\hat{\omega}_1 \pm z_{.975}\sqrt{\frac{\hat{\omega}_1(1-\hat{\omega}_1)}{\tilde{n}}}$$

where $\hat{\omega}_1 = \frac{n\bar{Y}_n+2}{n+4}$ and $\tilde{n} = n+4$.

- Sometimes known as "adding 2 successes in 4 trials".
- $\hat{\omega}_1$ is a special case of $E[\Omega \mid \boldsymbol{Y}_n]$ where $\alpha = \beta = 2$. i.e. "shrinking" to a prior mean of $E[\Omega] = 1/2$

The Agresti-Coull Interval



see 4_AGinterval.R

Principles for Setting Priors

In this course we will mostly rely on these two principles:

- Uninformative: set priors to be uninformative, i.e.

$$f_\Omega(\omega) = 1I(\omega \in \mathcal{P})$$

  Okay in situations we'll discuss here, but in edge-cases this approach can lead to poor estimation. "Uninformative" in the sense that $f_\Omega(\omega)/f_\Omega(\omega') = 1$ for any $\omega, \omega' \in \mathcal{P}$
- Hierarchical Priors: Place a second layer of priors on hyperparamers:

$$\Omega \mid \Gamma \sim f_{\Omega|\Gamma}(\omega \mid \gamma)$$

$$\Gamma \sim f_\Gamma(\gamma)$$

  can encode dependences across parameters in ways that induce tailored shrinkage.

What have we learned?

1. Bayesian probability statements have direct interpretations on parameters.
2. The posterior is a "compromise" between prior and MLE.
3. Inference for transformations is automatic.
4. Priors induce shrinkage - leading to more stable estimates in small samples...
5. ...this can lead to good frequentist properties.
6. Bayesian inference can be subjective - different analysts may have different priors about ω.

The benefits of (1)-(5) for causal inference are immediate. (6) makes the Bayesian paradigm especially intuitive for causal sensitivity analyses.

Bayesian Posterior Sampling with Stan

Overview of Stan

In the last example, we were able to find the posterior directly and obtain draws $\{\omega^{(m)}\}_{m=1}^{M}$ easily using beta random number generators. In most realistic models used in practice, the posterior will not have a known form.

Stan is a probabilistic programming language (PPL). There are many PPLs (Greta, PyMC3, Nimble, JAGS, BUGS, ...). A PPL is a programming language for specifying probabilistic (Bayesian?) models.

- Provides syntax for specifying a likelihood.
- Provides syntax for specifying a prior.
- Returns a set of posterior draws $\{\omega^{(m)}\}_{m=1}^{M}$ using sophisticated MCMC strategies in the back-end.

We will not go over MCMC methods and posterior sampling.

Beta-Bernoulli Example: `.stan` file

For i.id. observations $i = 1, 2, \ldots, n$

$$\begin{aligned} Y_i \mid \Omega &\sim Ber(\omega) \\ \Omega &\sim Beta(\alpha, \beta) \end{aligned}$$

where $\alpha > 0$,$\beta > 0$ are specified constants.

Example `beta_bernoulli_model.stan` file:

———

```
data {
  // values here are passed from R
  int<lower=0> n;
  int<lower=0, upper=1> y[n];
  real<lower=0> alpha;
  real<lower=0> beta;
}

parameters {
  real<lower=0, upper=1> omega;
}

model {
  omega ~ beta(alpha, beta);
  y ~ bernoulli(omega);
}
```